Deep Learning - Week 8

1. What are the challenges associated with using the Tanh(x) activation function?

    (a) It is not zero centered

    (b) Computationally expensive

    (c) Non-differentiable at 0

    (d) Saturation

    Correct Answer: (b),(d)
    Solution: Tanh(x) is zero-centered but the problem of saturation still persists. It is computationally expensive to do this operation.

2. Which of the following problems makes training a neural network harder while using sigmoid as the activation function?

    (a) Not-continuous at 0

    (b) Not-differentiable at 0

    (c) Saturation

    (d) Computationally expensive

    Correct Answer: (c),(d)
    Solution: Sigmoid is computationally expensive due to the exponentiation process. They saturate easily and since their range is [0,1], weight update directions are limited.

3. Consider the Exponential ReLU (ELU) activation function, defined as:

$$f(x) = \begin{cases} x, & x > 0 \\ a(e^x - 1), & x \le 0 \end{cases}$$

    where $a \neq 0$. Which of the following statements is true?

    (a) The function is discontinuous at $x = 0$.

    (b) The function is non-differentiable at $x = 0$.

    (c) Exponential ReLU can produce negative values.

    (d) Exponential ReLU is computationally less expensive than ReLU.

    Correct Answer: (c)
    Solution:

    1. **Discontinuity at $x = 0$:**

    (a) Right-hand limit: $\lim_{x \to 0^+} f(x) = 0$.

    (b) Left-hand limit: $\lim_{x \to 0^-} a(e^x - 1) = a(1 - 1) = 0$.

    (c) Since both limits and $f(0)$ are equal, the function is continuous at $x = 0$.

2. **Non-differentiability at** $x = 0$:

(a) Right derivative: $\lim\limits_{x \to 0^+} f'(x) = 1$.

(b) Left derivative: $\lim\limits_{x \to 0^-} ae^x = a$.

(c) The function is differentiable at $x = 0$ only if $a = 1$.

(d) Since $a \neq 0$ but not necessarily 1, differentiability depends on a, making this statement inconclusive.

3. **Computational expense compared to ReLU**:

(a) ReLU uses $\max(0, x)$, which is a simple comparison.

(b) ELU involves an exponential operation, which is more computationally expensive.

4. **Possibility of negative values**:

(a) For $x < 0$,
$$f(x) = a(e^x - 1).$$

(b) Since $e^x - 1 < 0$ for $x < 0$, $f(x)$ is negative if $a > 0$.

4. We have observed that the sigmoid neuron has become saturated. What might be the possible output values at this neuron?

(a) 0.0666

(b) 0.589

(c) 0.9734

(d) 0.498

(e) 1

Imp Change here

Correct Answer: (a),(c),(e)
Solution: Since the neuron has saturated its output values are close to 0 or 1.

5. What is the gradient of the sigmoid function at saturation?

Correct Answer: 0
Solution: At saturation, the sigmoid function outputs 0 or 1, and its gradient becomes zero, causing vanishing gradients.

6. Which of the following are common issues caused by saturating neurons in deep networks?

(a) Vanishing gradients

(b) Slow convergence during training

(c) Overfitting

(d) Increased model complexity

Solution: Saturating neurons, especially in sigmoid activation functions, cause vanishing gradients, making it hard to propagate error signals back and slow down learning.

7. Given a neuron initialized with weights $w_1 = 0.9$, $w_2 = 1.7$, and inputs $x_1 = 0.4$, $x_2 = -0.7$, calculate the output of a ReLU neuron.

Solution: The weighted sum is $0.9 \times 0.4 + 1.7 \times (-0.7) = 0.36 - 1.19 = -0.83$. ReLU outputs the max of 0 and the input, so the result is $max(0, -0.83) = 0$.

8. Which of the following is incorrect with respect to the batch normalization process in neural networks?

   (a) We normalize the output produced at each layer before feeding it into the next layer

   (b) Batch normalization leads to a better initialization of weights.

   (c) Backpropagation can be used after batch normalization

   (d) Variance and mean are not learnable parameters.

Solution:

1. "We normalize the output produced at each layer before feeding it into the next layer."
Batch Normalization (BN) normalizes activations by adjusting them to have zero mean and unit variance before passing them to the next layer.
The formula for batch normalization is:

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

This helps stabilize learning and speeds up convergence.

2. "Batch normalization leads to a better initialization of weights."
BN helps mitigate issues like internal covariate shift, making the training less dependent on careful weight initialization.
It allows training with higher learning rates and stabilizes deep networks.

3. "Backpropagation can be used after batch normalization."
BN is differentiable, and gradients can flow through it during backpropagation.
During training, gradients are computed normally, taking into account the transformation applied by BN.

4. "Variance and mean are not learnable parameters." (Incorrect)
BN initially normalizes using batch statistics (mean μ and variance σ^2).
However, batch normalization introduces learnable parameters: - γ (scaling parameter) - β (shifting parameter)
These parameters allow the model to learn an optimal representation instead of always enforcing zero mean and unit variance.

9. Which of the following is an advantage of unsupervised pre-training in deep learning?

   (a) It helps in reducing overfitting

   (b) Pre-trained models converge faster

   (c) It requires fewer computational resources

   (d) It improves the accuracy of the model

Correct Answer: (a),(b),(d)
Solution: Unsupervised pre-training helps in reducing overfitting in deep neural networks by initializing the weights in a better way. This technique requires more computational resources than supervised learning, but it can improve the accuracy of the model. Additionally, the pre-trained model is shown to converge faster than non-pre-trained models

10. How can you tell if your network is suffering from the Dead ReLU problem?

   (a) The loss function is not decreasing during training

   (b) A large number of neurons have zero output

   (c) The accuracy of the network is not improving

   (d) The network is overfitting to the training data

Correct Answer: (b)
Solution: The Dead ReLU problem can be detected by checking the output of each neuron in the network. If a large number of neurons have zero output, then the network may be suffering from the Dead ReLU problem. This can indicate that the bias term is too high, causing a large number of dead neurons.